INTRODUCTION TO R AND R FOR GENETIC DATA ANALYSIS

MRC Epidemiology Unit

A presentation for the Genetic Meeting on November 9, 2006

Three topics

1. R and general statistics, what is R and why it is necessary to know about it.

Ihaka R, Gentleman R. *R:* a language for data analysis and graphics. *J Comp Graph Stat* 1996; 5: 299-314.

Nolan D, Speed TP. Teaching statistics theory through application. *Am Statist* 1999; 53: 370-75.

Horton NJ, Brown ER, Qian LJ. Use of R as a toolbox for mathematical statistics exploration. *Am Stat* 2004; 58: 343-57.

2. R and statistical genetics, what is extra?

Zhao JH, Tan Q. Integrated analysis of genetic data with R. Hum Genomics. 2006; 2:258-65

Zhao JH, Tan Q. Genetic dissection of complex traits in silico: approaches, problems and solutions. Current Bioinformatics 2006; 1:359-69

3. Minimum set of tips and examples of genetic association analysis

CRAN task views for genetics

genetics, haplo.stats, gap and SNPassoc

However, this presentation is largely based on previous talks at KCL, UCL and U Southern Denmark and U Copenhagen

R AND GENERAL STATISTICS

Definition of R

R is “GNU S”, a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc’.

The name is partly based on the (first) names of the first two R authors (Robert Gentleman and Ross Ihaka), and partly a play on the name of the Bell Labs language “S”. S is a very high level language and an environment for data analysis and graphics. In 1998, the Association for Computing Machinery (ACM) presented its Software System Award to John M. Chambers, the principal designer of S, for the S system, which has forever altered the way people analyze, visualize, and manipulate data. . . S is an elegant, widely accepted, and enduring software system, with conceptual integrity, thanks to the insight, taste, and effort of John Chambers.

The four coloured (Brown, Blue, White, Green) books on S and other books

Richard A. Becker and John M. Chambers (1984) S. An Interactive Environment for Data Analysis and Graphics. Monterey: Wadsworth and Brooks/Cole.

Richard A. Becker, John M. Chambers and Allan R. Wilks (1988), “The New S Language,” London: Chapman & Hall.

John M. Chambers and Trevor J. Hastie (1992), “Statistical Models in S,” London: Chapman & Hall.

John M. Chambers (1998), “Programming with Data,” New York: Springer,

and moreover,

P. Dalgaard (2002) Introductory Statistics with R. Springer: New York

J. Fox (2002) An R and S-PLUS Companion to Applied Regression. Sage Publications

P. Murrell (2005) R Graphics. Chapman & Hall/CRC

Examples of analysis: linear and linear logistic regression, GLM, glmm

The following books are also helpful,

Venables WN, Ripley BD (2002) Modern Applied Statistics with S. 4th Edition

Pinheiro JC, Bates DM (2000) Mixed-Effects Models in S and S-PLUS. Springer

CRAN task views (http://cran.r-project.org/src/contrib/Views/)

R AND STATISTICAL GENETICS

There is a list of over 300 computer software from http://linkage.rockefeller.edu.

An integrated analysis of genetic data has in-house facility for traditional analysis. It is an environment of statistical analysis that provides facility for database accessibility, graphics, mathematical/statistical routines, flexible programming language, ability to incorporate available codes, Internet connectivity.

All analyses can be done effectively and efficiently.

Examples

Zhao JH. Pedigree-drawing with R and graphviz. Bioinformatics 2006; 22:1013–1014
Zhao JH. Mixed-effects Cox models of alcohol dependence in extended families. BMC Genet 2005; **6**(Suppl 1):S127
Tan Q, Zhao J, Li S, Kruse TA, Christensen K. Demographic analysis of microarray gene expression data in the CEPH Utah families. GAW15

Microarray data analysis through http://www.bioconductor.org

R FOR GENETIC ASSOCIATION ANALYSIS

The core packages as specified in CRAN task views for genetics (description as of July 2005):

genetics. Classes and methods for handling genetic data. Includes classes to represent genotypes and haplotypes at single markers up to multiple markers on multiple chromosomes. Functions include allele frequencies, flagging homo/heterozygotes, flagging carriers of certain alleles, estimating and testing for Hardy–Weinberg disequilibrium, estimating and testing for linkage disequilibrium.

haplo.stats. A suite of S-PLUS/R routines for the analysis of indirectly measured haplotypes.16 The statistical methods assume that all subjects are unrelated and that haplotypes are ambiguous (due to unknown linkage phase of the genetic markers). The genetic markers are assumed to be co-dominant (ie one-to-one correspondence between their genotypes and their phenotypes), and the measurements of genetic markers are referred to as genotypes. The main functions in haplo.stats are: haplo.em, haplo.glm and haplo.score. The haplo.score function is an extension of an earlier function in the haplo.score package.

R/gap. An integrated package for genetic data analysis of both population and family data. It contains functions for sample size calculations of both population- and family-based designs, probability of familial disease aggregation, kinship calculation, some statistics in linkage analysis and association analysis involving one or more genetic markers, including haplotype analysis. The functions included are: hwe, hwe.hardy for Hardy–Weinberg equilibria involving SNPs and highly polymorphic microsatellite markers; s2k, gcontrol for singlelocus association analysis of polymorphic markers and genomic control; genecounting; gcp for haplotype analysis of all chromosomes and missing data and permutation tests; tbyt, kbyl for linkage disequilibrium statistics for SNPs and multiallelic markers; htr, hap.score for extracting haplotype information for haplotype trend regression analysis and regression incorporating covariates based on conditional regression, as implemented in the haplo.score package. For family data, it includes family plotting through graphviz (pedtodot), exact probability of familial clustering disease (pfc and pfc.sim), kinship calculation, involves genetic index of familiality (gif) and a simple kinship calculation (kin.morgan).

The following is example code using SNPassoc package for the study of type-2 diabetes involving 5,013 Ashkenazi and four UK populations using two-staged design and 4,570 SNPs across a 10Mb region of chromosome 20q.

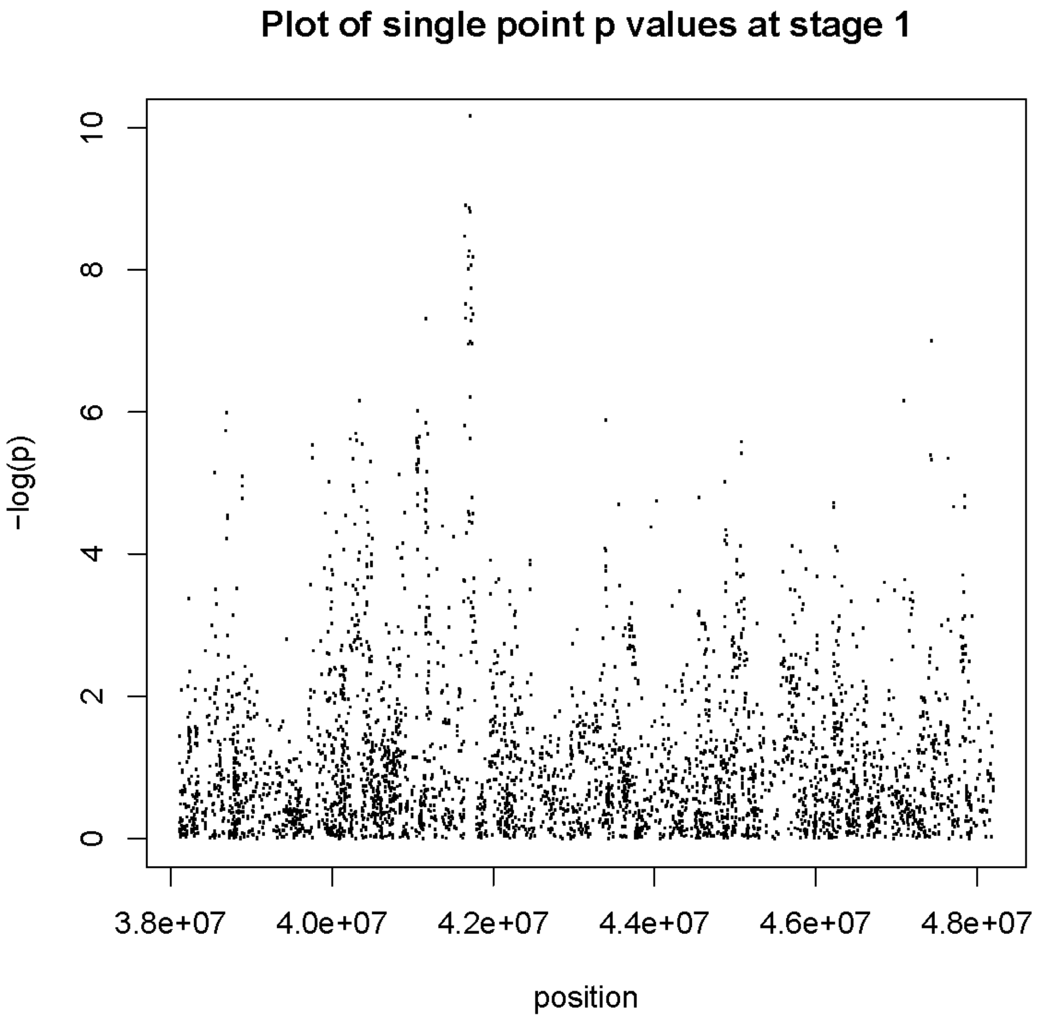
setup.R

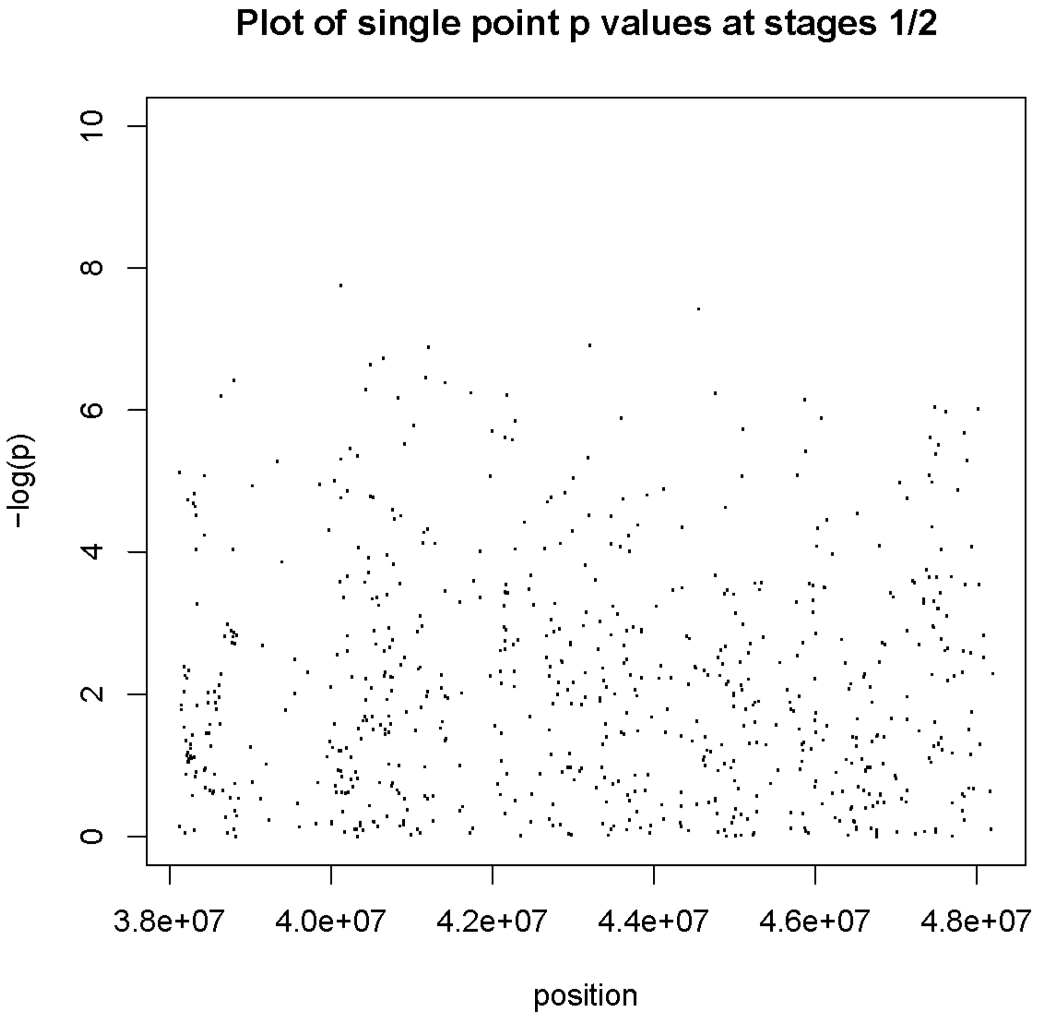
```
# 25/10/2006 MRC-Epid, JHZ

library(foreign)
g4 <- read.dta("g4.dta")
g4pos <- read.dta("g4pos.dta")
save(g4,g4pos,file="g4.Rdata")
library(SNPassoc)
snps <- setupSNP(g4, colSNPs=5:3961, sort = TRUE, info=g4pos)
save(snps,file="snps.Rdata")
q('no')
```

analyse.R

```
# 30/10/2006 MRC-Epid, JHZ

library(SNPassoc)
library(qvalue)
load("g4.Rdata")
load("snps.Rdata")

## single stage analysis
stage1 <- (snps$stage==1)
snps1 <- snps[stage1,]
ans1 <- WGassociation(cc~b58cregion, data=snps1, model="log")
psnps1 <- attr(ans1,"pvalues")
x <- attr(ans1,"gen.info")
pdf("stage1.pdf")
plot(x$pos,-log(psnps1[,2]),ylim=c(0,10),xlab="position",ylab="-
log(p)",type="p",pch=".")
title("Plot of single point p values at stage 1")
dev.off()

## joint analysis of SNPs at both stages
stage2 <- (snps$stage==2)
snps2sum <- apply(snps[stage2,-(1:4)],2,function(x) sum(is.na(x),na.rm=TRUE))
s1only <- (snps2sum==1978)
```

```
snps12 <- snps[,c(TRUE,TRUE,TRUE,TRUE,!s1only)]
snps12 <- setupSNP(snps12,colSNPs=5:dim(snps12)[2])
ans <- WGassociation(cc~stage+b58cregion, data=snps12, model="log")
x <- g4pos$pos[!s1only]
pdf("stage.pdf")
psnps <- attr(ans,"pvalues")
plot(x,-log(psnps[,2]),ylim=c(0,10),xlab="position",ylab="-log(p)",type="p",pch=".")
title("Plot of single point p values at stages 1/2")
qobj <- qvalue(psnps[,2])
qplot(qobj)
qwrite(qobj,filename="q.txt")
dev.off()
```

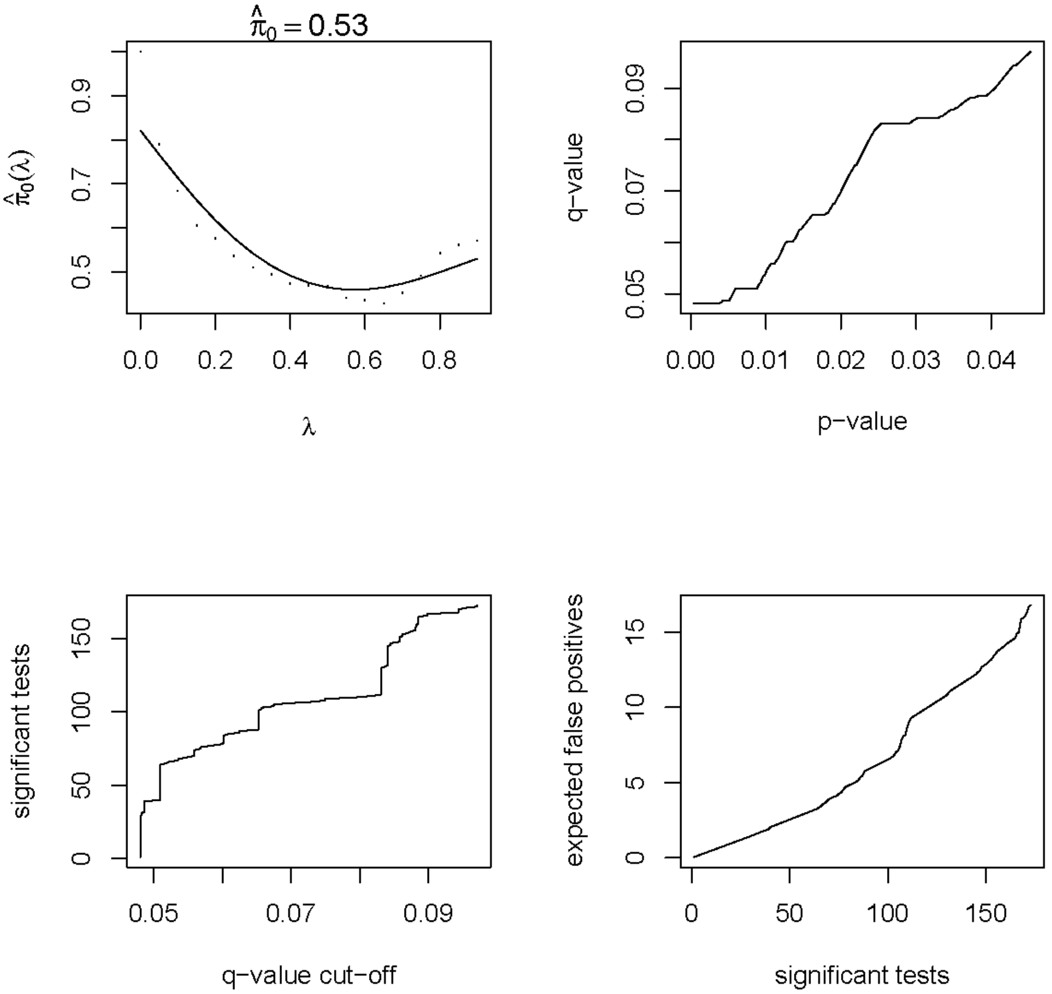




It is desirable to take into account of the correlation the statistics across stages.

The false discovery rate measures the proportion of false positives among all SNPs called significant: $FDR = \left[\frac{F}{S} \mid S > 0\right] p(S > 0)$ or $pFDR = E\left[\frac{F}{S} \mid S > 0\right]$, where S is the total number of tests called significant but only F of which are true. The p-value is a measure of significance in terms of false positive rate (Type I error rate). The *q*-value is an FDR measure of significance associate with each SNP. Since $q - value(p_i) = \min_{t \geq p_i} pFDR(t)$, $\forall t \in [0,1]$ and for a large number of SNPs, $P(S > 0) = 1$ we can use FDR for the estimate of *q*-value. The λ parameter is chosen to tune for the proportion of truly null (π_0) to obtain the FDR estimate or approximately *q*-value.

There are other packages such as locfdr and fdr-library available. The package multtest contains many other procedures.

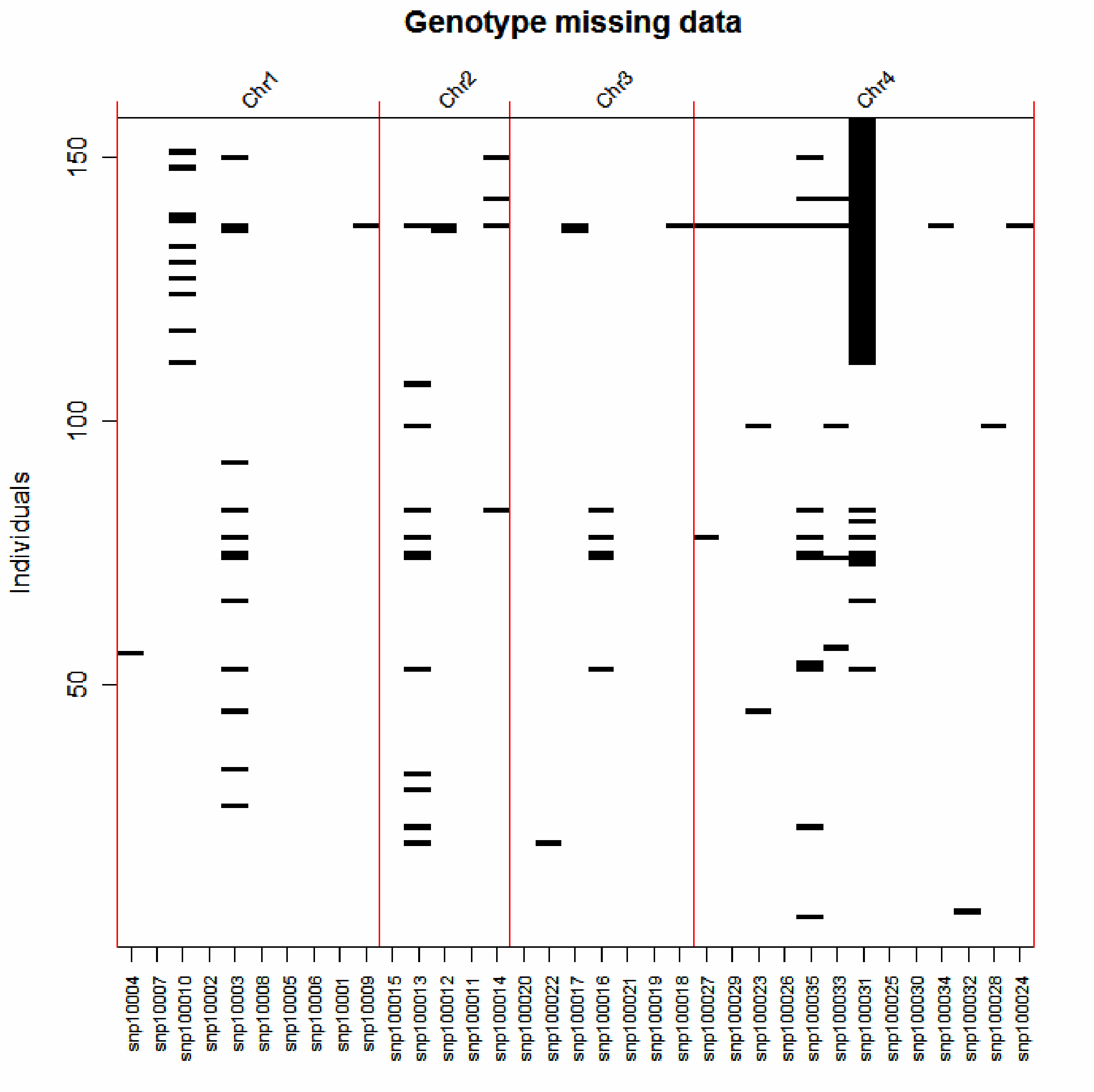


Documentation example

tableHWE

```
         all.groups   Male Female
snp10001      0.2816 0.3941 0.7388
snp10002      0.0049 0.1660 0.0075
snp10003           -      -      -
snp10004           -      -      -
snp10005      0.0080 0.2755 0.0257
snp10006           -      -      -
snp10007           -      -      -
snp10008      0.1378 0.5078 0.2342
…
```

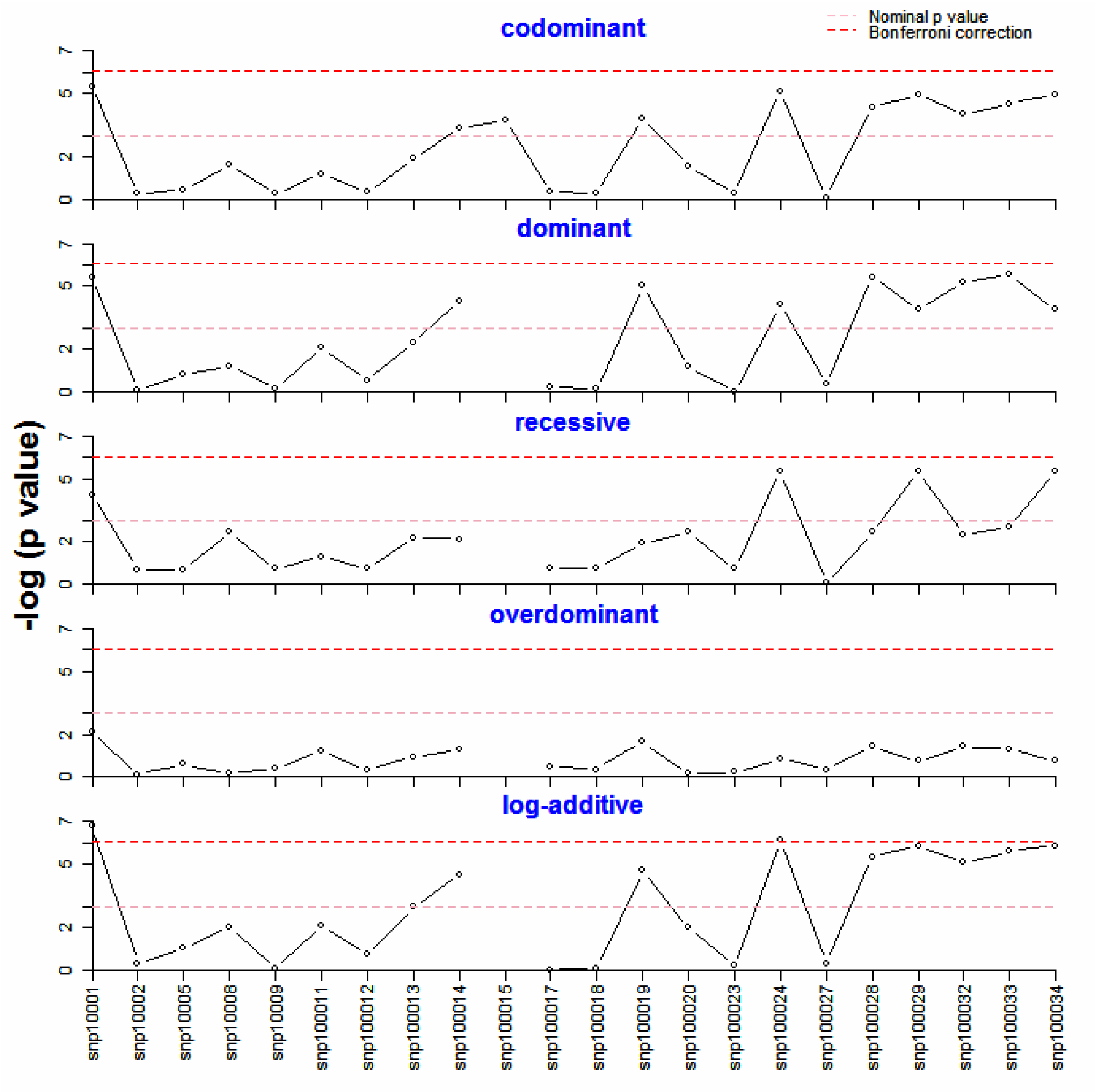
plotMissing



association

```
SNP: snp10001  adjusted by:
                 0    %   1    %   OR lower upper p-value   AIC
Codominant
T/T             24 51.1  68 61.8 1.00                0.1323 193.6
C/T             21 44.7  32 29.1 0.54  0.26  1.11
C/C              2  4.3  10  9.1 1.76  0.36  8.64
Dominant
T/T             24 51.1  68 61.8 1.00                0.2118 194.1
C/T-C/C         23 48.9  42 38.2 0.64  0.32  1.28
Recessive
T/T-C/T         45 95.7 100 90.9 1.00                0.2715 194.4
C/C              2  4.3  10  9.1 2.25  0.47 10.69
Overdominant
T/T             26 55.3  78 70.9 1.00                0.0613 192.1
T/T-C/C         21 44.7  32 29.1 0.51  0.25  1.03
log-Additive
0,1,2           47 29.9 110 70.1 0.87  0.51  1.47   0.5945 195.4
```

WGassociation



Bonferoni.sig

```
number of tests:  22
alpha: 0.05
corrected alpha: 0.002272727
   No significant SNPs after Bonferroni correction
```

SUMMARY

In general, R offers a wide range of packages useful for epidemiological analyses as well as candidate gene analyses.

The long-term benefit will surely outweigh the effort to climb the learning curve

Date created 31-10-2006